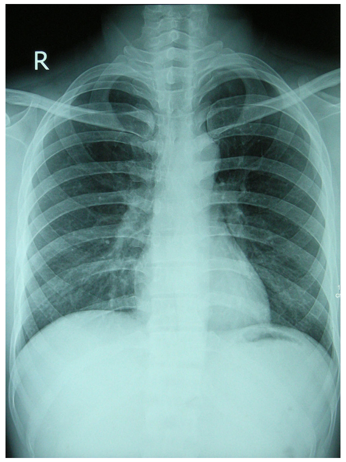
Toward a taxonomy of trust for probabilistic data analysis

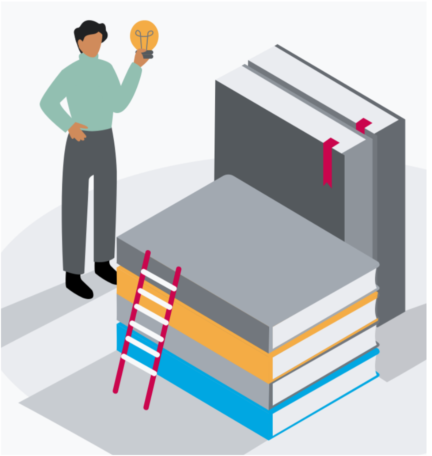
Tamara Broderick
Associate Professor,
MIT

With Ryan Giordano, Andrew Gelman, Rachael Meager,
Anna L. Smith, and Tian Zheng

When can I trust decisions from data?

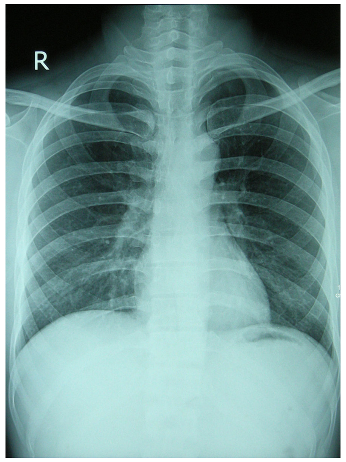
- More data & better computation ➔ data analyses increasingly drive life-changing decisions

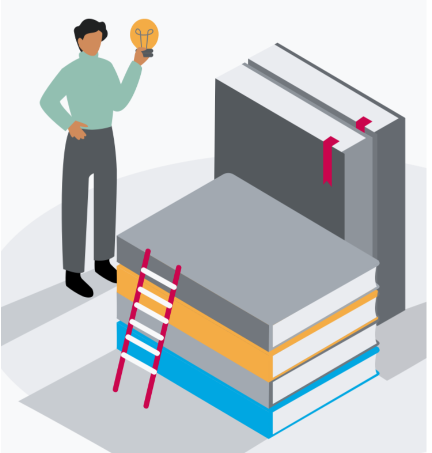




When can I trust decisions from data?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions

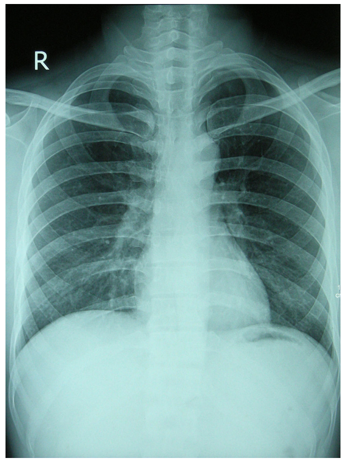


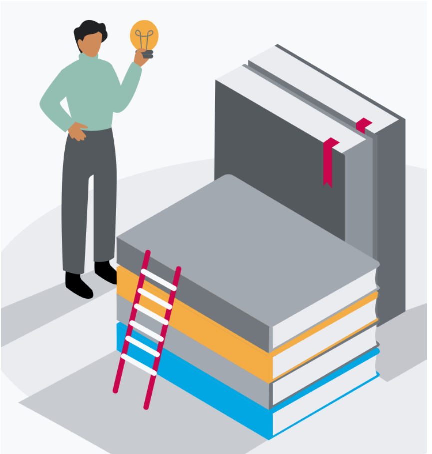


- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
  - E.g. (1) conclude microcredit helps, then (2) distribute it

When can I trust decisions from data?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions

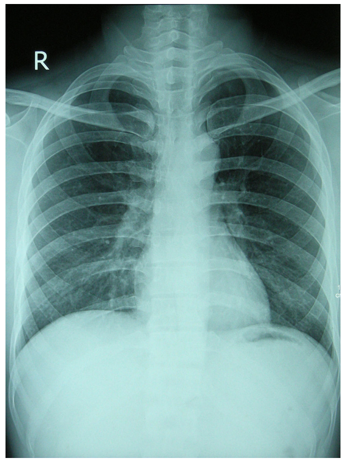


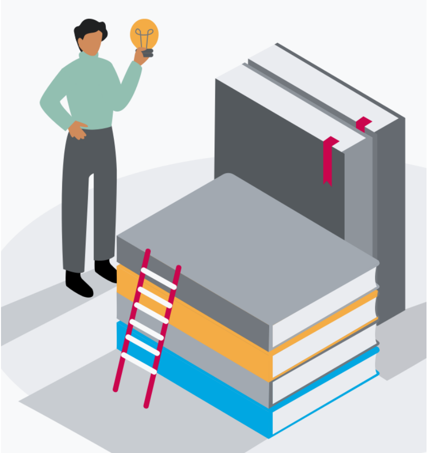


- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
  - E.g. (1) conclude microcredit helps, then (2) distribute it
- Might worry about *generalization* if replicability fails:

When can I trust decisions from data?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions



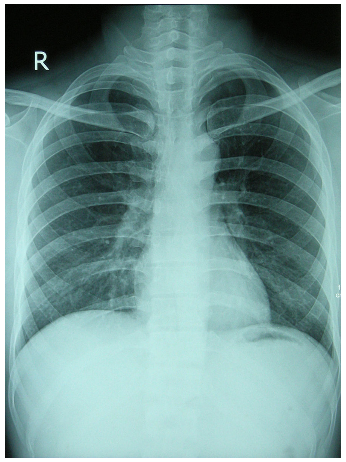
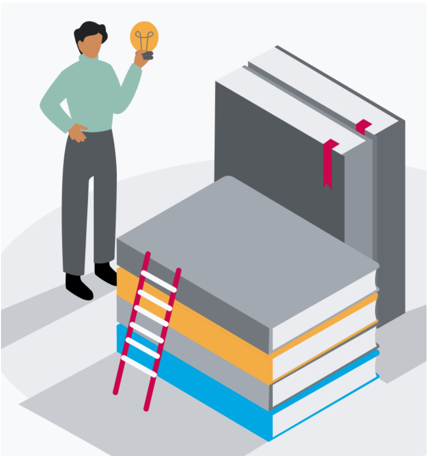


- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
  - E.g. (1) conclude microcredit helps, then (2) distribute it
- Might worry about *generalization* if replicability fails: E.g. in repeat of 100 psych experiments, <1/2 had same result

[Open Science Collaboration, 2015]

When can I trust decisions from data?

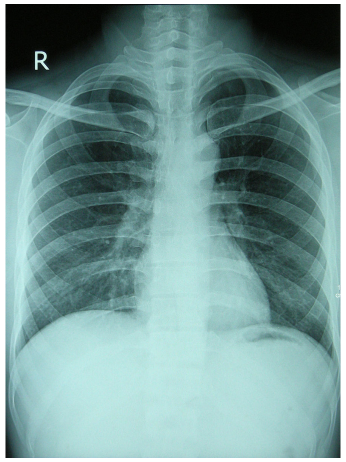
- More data & better computation ➔ data analyses increasingly drive life-changing decisions

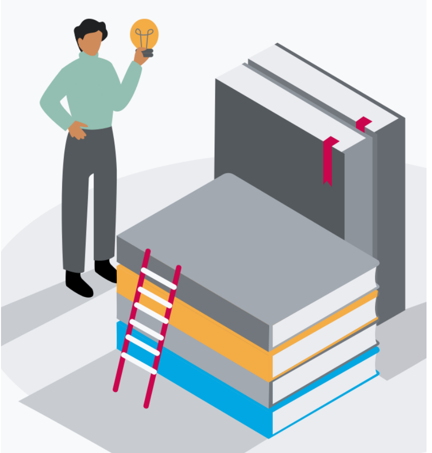




- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
  - E.g. (1) conclude microcredit helps, then (2) distribute it
- Might worry about *generalization* if replicability fails: E.g. in repeat of 100 psych experiments, <1/2 had same result [Open Science Collaboration, 2015]
  - Of 53 hematology/oncology papers, 6 had same result [Begley, Ellis 2012]

When can I trust decisions from data?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions

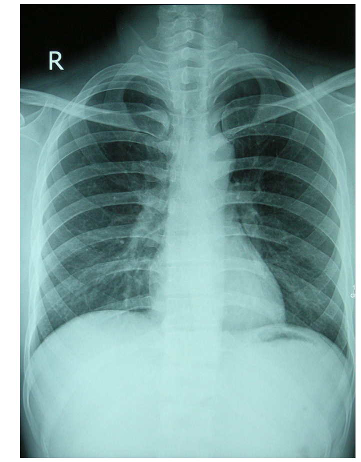



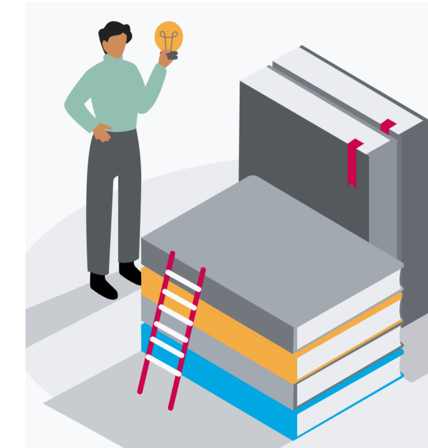


- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
  - E.g. (1) conclude microcredit helps, then (2) distribute it
- Might worry about *generalization* if replicability fails: E.g. in repeat of 100 psych experiments, <1/2 had same result [Open Science Collaboration, 2015]
  - Of 53 hematology/oncology papers, 6 had same result [Begley, Ellis 2012]
- **We review**: how generalization could fail (when everyone is well-meaning & using vetted tools) and mitigations

When can I trust decisions from data?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions





- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
  - E.g. (1) conclude microcredit helps, then (2) distribute it
- Might worry about *generalization* if replicability fails: E.g. in repeat of 100 psych experiments, <1/2 had same result
  [Open Science Collaboration, 2015]
  - Of 53 hematology/oncology papers, 6 had same result
    [Begley, Ellis 2012]
- **We review**: how generalization could fail (when everyone is well-meaning & using vetted tools) and mitigations
  - And **we propose** a check for stability (not a cure-all)

Roadmap

Roadmap

- An example analysis: microcredit

Roadmap

- An example analysis: microcredit
  - Collecting data
  - Turning your real-life problem into math
  - Algorithms
  - Code

Roadmap

- An example analysis: microcredit
- Challenges and mitigations
  - Collecting data
  - Turning your real-life problem into math
  - Algorithms
  - Code

Roadmap

- An example analysis: microcredit
- Challenges and mitigations
  - Collecting data
  - Turning your real-life problem into math
  - Algorithms
  - Code
- A stability check: can dropping a small fraction of data change conclusions?

Roadmap

- An example analysis: microcredit
- Challenges and mitigations
    - Collecting data
    - Turning your real-life problem into math
    - Algorithms
    - Code
- A stability check: can dropping a small fraction of data change conclusions?
    - In an existing study, we can drop 1 out of >16,500 data points and flip the sign of the effect

Roadmap

- An example analysis: microcredit
- Challenges and mitigations
  - Collecting data
  - Turning your real-life problem into math
  - Algorithms
  - Code
- A stability check: can dropping a small fraction of data change conclusions?
  - In an existing study, we can drop 1 out of >16,500 data points and flip the sign of the effect
  - This check isn't equivalent to standard tools (gross outliers, p-values, etc)

Roadmap

- An example analysis: microcredit
- Challenges and mitigations
  - Collecting data
  - Turning your real-life problem into math
  - Algorithms
  - Code
- A stability check: can dropping a small fraction of data change conclusions?
  - In an existing study, we can drop 1 out of >16,500 data points and flip the sign of the effect
  - This check isn't equivalent to standard tools (gross outliers, p-values, etc)

Example analysis: microcredit

Example analysis: microcredit

Real-world goal: Want to know if microcredit helps people

Example analysis: microcredit

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
    - What counts as “help”? E.g. increase business profit

Example analysis: microcredit

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc

Example analysis: microcredit

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc
   - How is data collected? Survey: In-person, phone, mail

Example analysis: microcredit

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
    - What counts as “help”? E.g. increase business profit
    - Who gets measured? Particular people, location, etc
    - How is data collected? Survey: In-person, phone, mail
    - How is it distributed? Randomized controlled trial

Example analysis: microcredit

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
    - What counts as “help”? E.g. increase business profit
    - Who gets measured? Particular people, location, etc
    - How is data collected? Survey: In-person, phone, mail
    - How is it distributed? Randomized controlled trial
2. Have to figure out how to make a decision from this data

Example analysis: microcredit

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc
   - How is data collected? Survey: In-person, phone, mail
   - How is it distributed? Randomized controlled trial
2. Have to figure out how to make a decision from this data
   - What counts as “increasing profit”? E.g. Compare mean profit of groups receiving and not receiving microcredit. Look for a statistically significant positive effect.

Example analysis: microcredit

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
    - What counts as “help”? E.g. increase business profit
    - Who gets measured? Particular people, location, etc
    - How is data collected? Survey: In-person, phone, mail
    - How is it distributed? Randomized controlled trial
2. Have to figure out how to make a decision from this data
    - What counts as “increasing profit”? E.g. Compare mean profit of groups receiving and not receiving microcredit. Look for a statistically significant positive effect.
3. Have to choose an algorithm: E.g. ordinary least squares, Markov chain Monte Carlo, a deep learning method

Example analysis: microcredit

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
    - What counts as “help”? E.g. increase business profit
    - Who gets measured? Particular people, location, etc
    - How is data collected? Survey: In-person, phone, mail
    - How is it distributed? Randomized controlled trial
2. Have to figure out how to make a decision from this data
    - What counts as “increasing profit”? E.g. Compare mean profit of groups receiving and not receiving microcredit. Look for a statistically significant positive effect.
3. Have to choose an algorithm: E.g. ordinary least squares, Markov chain Monte Carlo, a deep learning method
4. Have to choose code: Packages, but also full pipeline

Example analysis: microcredit

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into **measurements on actual data**
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc
   - How is data collected? Survey: In-person, phone, mail
   - How is it distributed? Randomized controlled trial
2. Have to figure out how to **make a decision** from this data
   - What counts as “increasing profit”? E.g. Compare mean profit of groups receiving and not receiving microcredit. Look for a statistically significant positive effect.
3. Have to choose an **algorithm**: E.g. ordinary least squares, Markov chain Monte Carlo, a deep learning method
4. Have to choose **code**: Packages, but also full pipeline

Choosing actual measurements

- It’s hard to measure what I care about. Is a proxy ok?

Choosing actual measurements

- It’s hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let’s hope so! But:

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia
- If I run a randomized controlled trial, doesn't that guarantee any benefit I find will generalize? Job placement assistance

[Crépon et al 2013]

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia
- If I run a randomized controlled trial, doesn't that guarantee any benefit I find will generalize? Job placement assistance

[Crépon et al 2013]

  - Concluded benefits at 8 months, gone by 12 months

Choosing actual measurements

- It’s hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let’s hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia
- If I run a randomized controlled trial, doesn’t that guarantee any benefit I find will generalize? Job placement assistance

[Crépon et al 2013]

  - Concluded benefits at 8 months, gone by 12 months
  - Benefits seemed to be at expense of other workers

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia
- If I run a randomized controlled trial, doesn't that guarantee any benefit I find will generalize? Job placement assistance

[Crépon et al 2013]

  - Concluded benefits at 8 months, gone by 12 months
  - Benefits seemed to be at expense of other workers
- Mitigations: domain expertise, context, team science

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia
- If I run a randomized controlled trial, doesn't that guarantee any benefit I find will generalize? Job placement assistance

[Crépon et al 2013]

  - Concluded benefits at 8 months, gone by 12 months
  - Benefits seemed to be at expense of other workers
- Mitigations: domain expertise, context, team science
  - Importance of longer-range/larger-scale experiments (≠many small experiments), incentives, funding

Code: challenges & mitigations

- Boring! I know about bugs already. Why even talk about it?

Code: challenges & mitigations

- Boring! I know about bugs already. Why even talk about it?
  - An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error

[Reinhart, Rogoff 2010; Herndon et al 2014]

Code: challenges & mitigations

- Boring! I know about bugs already. Why even talk about it?
  - An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error [Reinhart, Rogoff 2010; Herndon et al 2014]
  - Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials [Baggerly, Coombes 2009]

Code: challenges & mitigations

- Boring! I know about bugs already. Why even talk about it?
  - An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error [Reinhart, Rogoff 2010; Herndon et al 2014]
  - Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials [Baggerly, Coombes 2009]
- OK but I use standard packages and data analyses, so why would I ever worry about bugs?

Code: challenges & mitigations

- Boring! I know about bugs already. Why even talk about it?
  - An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error [Reinhart, Rogoff 2010; Herndon et al 2014]
  - Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials [Baggerly, Coombes 2009]
- OK but I use standard packages and data analyses, so why would I ever worry about bugs?
  - Natural to outsource code (team science)

Code: challenges & mitigations

- Boring! I know about bugs already. Why even talk about it?
  - An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error [Reinhart, Rogoff 2010; Herndon et al 2014]
  - Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials [Baggerly, Coombes 2009]
- OK but I use standard packages and data analyses, so why would I ever worry about bugs?
  - Natural to outsource code (team science)
  - The issues highlighted above are in pre-processing.

Code: challenges & mitigations

- Boring! I know about bugs already. Why even talk about it?
  - An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error
    [Reinhart, Rogoff 2010; Herndon et al 2014]
  - Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials
    [Baggerly, Coombes 2009]
- OK but I use standard packages and data analyses, so why would I ever worry about bugs?
  - Natural to outsource code (team science)
  - The issues highlighted above are in pre-processing.
- That scientific area over there has a problem with bugs (or replicability or another issue), not mine.

Code: challenges & mitigations

- Boring! I know about bugs already. Why even talk about it?
  - An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error [Reinhart, Rogoff 2010; Herndon et al 2014]
  - Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials [Baggerly, Coombes 2009]
- OK but I use standard packages and data analyses, so why would I ever worry about bugs?
  - Natural to outsource code (team science)
  - The issues highlighted above are in pre-processing.
- That scientific area over there has a problem with bugs (or replicability or another issue), not mine.
  - Biased sample: We find more problems when people check for problems. (Easy to disincentivize checking.)

Code: challenges & mitigations

- So what can we do?

Code: challenges & mitigations

- So what can we do?
  - Bugs aren't solved by spreading awareness, even of particular bugs. 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates; 2021: 30% of studied papers had this issue

[Zeeberg et al 2004; Zieman et al 2016; Abeysooriya et al 2021; Lewis 2021]

Code: challenges & mitigations

- So what can we do?
  - Bugs aren't solved by spreading awareness, even of particular bugs. 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates; 2021: 30% of studied papers had this issue
    [Zeeberg et al 2004; Zieman et al 2016; Abeysooriya et al 2021; Lewis 2021]
  - Shaming people doesn't seem very effective. People worry their code isn't perfect (it's not!) and don't share it.

Code: challenges & mitigations

- So what can we do?
    - Bugs aren’t solved by spreading awareness, even of particular bugs. 2016: 20% of studied papers had Excel error that “autocorrects” certain gene names into dates; 2021: 30% of studied papers had this issue
      [Zeeberg et al 2004; Zieman et al 2016; Abeysooriya et al 2021; Lewis 2021]
    - Shaming people doesn’t seem very effective. People worry their code isn’t perfect (it’s not!) and don’t share it.
    - Start: journals require code. Next: enforcement, checks.

Code: challenges & mitigations

- So what can we do?
  - Bugs aren't solved by spreading awareness, even of particular bugs. 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates; 2021: 30% of studied papers had this issue
    [Zeeberg et al 2004; Zieman et al 2016; Abeysooriya et al 2021; Lewis 2021]
  - Shaming people doesn't seem very effective. People worry their code isn't perfect (it's not!) and don't share it.
  - Start: journals require code. Next: enforcement, checks.
  - Tools to support tracking models/packages/pipelines.
    [Stodden et al 2014, 2016; Vartak et al 2016; Gebru et al 2021; Heil et al 2021; wandb.ai]

Code: challenges & mitigations

- So what can we do?
    - Bugs aren’t solved by spreading awareness, even of particular bugs. 2016: 20% of studied papers had Excel error that “autocorrects” certain gene names into dates; 2021: 30% of studied papers had this issue
      [Zeeberg et al 2004; Zieman et al 2016; Abeysooriya et al 2021; Lewis 2021]
    - Shaming people doesn’t seem very effective. People worry their code isn’t perfect (it’s not!) and don’t share it.
    - Start: journals require code. Next: enforcement, checks.
    - Tools to support tracking models/packages/pipelines.
      [Stodden et al 2014, 2016; Vartak et al 2016; Gebru et al 2021; Heil et al 2021; wandb.ai]
- But this code is complex. Does it really need to be shared?

Code: challenges & mitigations

- So what can we do?
  - Bugs aren't solved by spreading awareness, even of particular bugs. 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates; 2021: 30% of studied papers had this issue
    [Zeeberg et al 2004; Zieman et al 2016; Abeysooriya et al 2021; Lewis 2021]
  - Shaming people doesn't seem very effective. People worry their code isn't perfect (it's not!) and don't share it.
  - Start: journals require code. Next: enforcement, checks.
  - Tools to support tracking models/packages/pipelines.
    [Stodden et al 2014, 2016; Vartak et al 2016; Gebru et al 2021; Heil et al 2021; wandb.ai]
- But this code is complex. Does it really need to be shared?
  - *Nature* gives pass to AI for detecting breast cancer: no code or full detail of the method
    [McKinney et al 2020; Haibe-Kains 2021]

Code: challenges & mitigations

- So what can we do?
  - Bugs aren't solved by spreading awareness, even of particular bugs. 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates; 2021: 30% of studied papers had this issue
    [Zeeberg et al 2004; Zieman et al 2016; Abeysooriya et al 2021; Lewis 2021]
  - Shaming people doesn't seem very effective. People worry their code isn't perfect (it's not!) and don't share it.
  - Start: journals require code. Next: enforcement, checks.
  - Tools to support tracking models/packages/pipelines.
    [Stodden et al 2014, 2016; Vartak et al 2016; Gebru et al 2021; Heil et al 2021; wandb.ai]
- But this code is complex. Does it really need to be shared?
  - *Nature* gives pass to AI for detecting breast cancer: no code or full detail of the method [McKinney et al 2020; Haibe-Kains 2021]
  - 13/62 AI methods for diagnosis from images gave code
    [Roberts et al 2021]

Code: challenges & mitigations

- So what can we do?
  - Bugs aren't solved by spreading awareness, even of particular bugs. 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates; 2021: 30% of studied papers had this issue
    [Zeeberg et al 2004; Zieman et al 2016; Abeysooriya et al 2021; Lewis 2021]
  - Shaming people doesn't seem very effective. People worry their code isn't perfect (it's not!) and don't share it.
  - Start: journals require code. Next: enforcement, checks.
  - Tools to support tracking models/packages/pipelines.
    [Stodden et al 2014, 2016; Vartak et al 2016; Gebru et al 2021; Heil et al 2021; wandb.ai]
- But this code is complex. Does it really need to be shared?
  - *Nature* gives pass to AI for detecting breast cancer: no code or full detail of the method [McKinney et al 2020; Haibe-Kains 2021]
  - 13/62 AI methods for diagnosis from images gave code [Roberts et al 2021]
  - Other groups can't check (and can't build on it)

Algorithms: a brief discussion

Algorithms: a brief discussion

- Didn’t the machine learners/statisticians prove this algorithm works? If there’s an issue, isn’t their problem?

Algorithms: a brief discussion

- Didn't the machine learners/statisticians prove this algorithm works? If there's an issue, isn't their problem?
  - Again, natural to outsource some development

Algorithms: a brief discussion

- Didn’t the machine learners/statisticians prove this algorithm works? If there’s an issue, isn’t their problem?
  - Again, natural to outsource some development
  - Can make simulated data to check how it behaves

Algorithms: a brief discussion

- Didn’t the machine learners/statisticians prove this algorithm works? If there’s an issue, isn’t their problem?
  - Again, natural to outsource some development
  - Can make simulated data to check how it behaves
  - Besides code, important to be aware that algorithms are typically justified under certain assumptions — and any guarantees are mathematical

Algorithms: a brief discussion

- Didn't the machine learners/statisticians prove this algorithm works? If there's an issue, isn't their problem?
  - Again, natural to outsource some development
  - Can make simulated data to check how it behaves
  - Besides code, important to be aware that algorithms are typically justified under certain assumptions — and any guarantees are mathematical

Turning high-level goals into math

- OK but if we're good data analysts with bug-free code and the same data, we'll all reach the same conclusions, right?

Algorithms: a brief discussion

- Didn’t the machine learners/statisticians prove this algorithm works? If there’s an issue, isn’t their problem?
  - Again, natural to outsource some development
  - Can make simulated data to check how it behaves
  - Besides code, important to be aware that algorithms are typically justified under certain assumptions — and any guarantees are mathematical

Turning high-level goals into math

- OK but if we’re good data analysts with bug-free code and the same data, we’ll all reach the same conclusions, right?
  - 29 teams used same data to answer “are soccer referees more likely to give red cards to dark-skin-toned players?”

[Silberzahn et al 2018]

Algorithms: a brief discussion

- Didn’t the machine learners/statisticians prove this algorithm works? If there’s an issue, isn’t their problem?
  - Again, natural to outsource some development
  - Can make simulated data to check how it behaves
  - Besides code, important to be aware that algorithms are typically justified under certain assumptions — and any guarantees are mathematical

Turning high-level goals into math

- OK but if we’re good data analysts with bug-free code and the same data, we’ll all reach the same conclusions, right?
  - 29 teams used same data to answer “are soccer referees more likely to give red cards to dark-skin-toned players?” [Silberzahn et al 2018]
    - 20 teams concluded yes, 9 teams did not

Algorithms: a brief discussion

- Didn't the machine learners/statisticians prove this algorithm works? If there's an issue, isn't their problem?
  - Again, natural to outsource some development
  - Can make simulated data to check how it behaves
  - Besides code, important to be aware that algorithms are typically justified under certain assumptions — and any guarantees are mathematical

Turning high-level goals into math

- OK but if we're good data analysts with bug-free code and the same data, we'll all reach the same conclusions, right?
  - 29 teams used same data to answer "are soccer referees more likely to give red cards to dark-skin-toned players?" [Silberzahn et al 2018]
    - 20 teams concluded yes, 9 teams did not
    - Did not find variability due to beliefs, expertise, quality

Turning high-level goals into math

- It’s hard to turn high-level goals into math

Turning high-level goals into math

- It’s hard to turn high-level goals into math
- High-level goal: Diagnose disease

Turning high-level goals into math

- It’s hard to turn high-level goals into math
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data

Turning high-level goals into math

- It’s hard to turn high-level goals into math
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data
  - AI for detecting skin cancer can learn to recognize experts’ ink marks

[Winkler et al 2019]

Turning high-level goals into math

- It’s hard to turn high-level goals into math
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data
    - AI for detecting skin cancer can learn to recognize experts’ ink marks [Winkler et al 2019]
    - AI using chest scans of children as no-covid controls can learn to identify children rather than lack of covid [Roberts et al 2021; Heaven 2021]

Turning high-level goals into math

- It's hard to turn high-level goals into math
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data
  - AI for detecting skin cancer can learn to recognize experts' ink marks [Winkler et al 2019]
  - AI using chest scans of children as no-covid controls can learn to identify children rather than lack of covid [Roberts et al 2021; Heaven 2021]
  - If full dataset has duplicates (e.g. in amalgam datasets), AI can learn to identify the particular patient or scan [Roberts et al 2021; Heaven 2021]

Turning high-level goals into math

- It's hard to turn high-level goals into math
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data
    - AI for detecting skin cancer can learn to recognize experts' ink marks [Winkler et al 2019]
    - AI using chest scans of children as no-covid controls can learn to identify children rather than lack of covid [Roberts et al 2021; Heaven 2021]
    - If full dataset has duplicates (e.g. in amalgam datasets), AI can learn to identify the particular patient or scan [Roberts et al 2021; Heaven 2021]
- Optimizing this particular objective is a convenient proxy, but it isn't the same as diagnosing disease well

Turning high-level goals into math

- It's hard to turn high-level goals into math
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data
  - AI for detecting skin cancer can learn to recognize experts' ink marks [Winkler et al 2019]
  - AI using chest scans of children as no-covid controls can learn to identify children rather than lack of covid [Roberts et al 2021; Heaven 2021]
  - If full dataset has duplicates (e.g. in amalgam datasets), AI can learn to identify the particular patient or scan [Roberts et al 2021; Heaven 2021]
- Optimizing this particular objective is a convenient proxy, but it isn't the same as diagnosing disease well
- Why cross-validation isn't a cure-all

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Formalization: Decide microcredit is helpful if mean business profit is higher in groups receiving microcredit

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Formalization: Decide microcredit is helpful if mean business profit is higher in groups receiving microcredit
  - Imagine a world where no one benefits from microcredit except for a tiny handful of people → Conclude: it helps

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Formalization: Decide microcredit is helpful if mean business profit is higher in groups receiving microcredit
  - Imagine a world where no one benefits from microcredit except for a tiny handful of people → Conclude: it helps
  - In US, household net worth mean ~$1M, median ~$193K

[Federal Reserve Board's Division of Research and Statistics, 2023]

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Formalization: Decide microcredit is helpful if mean business profit is higher in groups receiving microcredit
  - Imagine a world where no one benefits from microcredit except for a tiny handful of people → Conclude: it helps
  - In US, household net worth mean ~$1M, median ~$193K [Federal Reserve Board's Division of Research and Statistics, 2023]
  - Linear models + ordinary least squares are standard in many areas of science & social science
    - (Typically) closed-form, unique solution. Well-vetted code, theory. Relatively easy to understand

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Formalization: Decide microcredit is helpful if mean business profit is higher in groups receiving microcredit
  - Imagine a world where no one benefits from microcredit except for a tiny handful of people → Conclude: it helps
  - In US, household net worth mean ~$1M, median ~$193K [Federal Reserve Board's Division of Research and Statistics, 2023]
  - Linear models + ordinary least squares are standard in many areas of science & social science
    - (Typically) closed-form, unique solution. Well-vetted code, theory. Relatively easy to understand
- Removing outliers isn't a panacea: E.g. ozone depletion first flagged as outliers to NASA (then checked)

[Earth Observatory, NASA, 2001; Pukelsheim, 1990]

Turning high-level goals into math

- But don’t p-values tell me if my result is right/generalizable?

Turning high-level goals into math

- But don't p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one model; if it's unlikely, we "reject" that model

Turning high-level goals into math

- But don’t p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one model; if it’s unlikely, we “reject” that model
  - “All models are wrong,” so expect p-value to get small with enough data (but that’s not necessarily meaningful)

[Box, 1976; Wang, Long 2022]

Turning high-level goals into math

- But don't p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one model; if it's unlikely, we "reject" that model
  - "All models are wrong," so expect p-value to get small with enough data (but that's not necessarily meaningful)

[Box, 1976; Wang, Long 2022]

- So what can we do?

Turning high-level goals into math

- But don't p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one model; if it's unlikely, we "reject" that model
  - "All models are wrong," so expect p-value to get small with enough data (but that's not necessarily meaningful)

[Box, 1976; Wang, Long 2022]

- So what can we do?
  - Bin Yu advocates for the importance of stability

[Yu, 2013, 2020; Yu, Kumbier 2020]

Turning high-level goals into math

- But don't p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one model; if it's unlikely, we "reject" that model
  - "All models are wrong," so expect p-value to get small with enough data (but that's not necessarily meaningful)

[Box, 1976; Wang, Long 2022]

- So what can we do?
  - Bin Yu advocates for the importance of stability

[Yu, 2013, 2020; Yu, Kumbier 2020]

  - Microcredit: multiple randomized controlled trials in different countries by different researchers

Turning high-level goals into math

- But don't p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one model; if it's unlikely, we "reject" that model
  - "All models are wrong," so expect p-value to get small with enough data (but that's not necessarily meaningful)

[Box, 1976; Wang, Long 2022]

- So what can we do?
  - Bin Yu advocates for the importance of stability

    [Yu, 2013, 2020; Yu, Kumbier 2020]
  - Microcredit: multiple randomized controlled trials in different countries by different researchers
  - Importance of incentivizing follow-up work, replication

Turning high-level goals into math

- But don't p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one model; if it's unlikely, we "reject" that model
  - "All models are wrong," so expect p-value to get small with enough data (but that's not necessarily meaningful)

[Box, 1976; Wang, Long 2022]

- So what can we do?
  - Bin Yu advocates for the importance of stability

[Yu, 2013, 2020; Yu, Kumbier 2020]

  - Microcredit: multiple randomized controlled trials in different countries by different researchers
  - Importance of incentivizing follow-up work, replication
  - Not always feasible to convene multiple teams

Turning high-level goals into math

- But don't p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one model; if it's unlikely, we "reject" that model
  - "All models are wrong," so expect p-value to get small with enough data (but that's not necessarily meaningful)

[Box, 1976; Wang, Long 2022]

- So what can we do?
  - Bin Yu advocates for the importance of stability

[Yu, 2013, 2020; Yu, Kumbier 2020]

  - Microcredit: multiple randomized controlled trials in different countries by different researchers
  - Importance of incentivizing follow-up work, replication
  - Not always feasible to convene multiple teams
  - Explainability as a form of stability

[Arrieta et al 2020; Doshi-Velez et al 2017; Zhang et al 2020; Mittelstadt et al 2019]

Turning high-level goals into math

- But don't p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one model; if it's unlikely, we "reject" that model
  - "All models are wrong," so expect p-value to get small with enough data (but that's not necessarily meaningful)

[Box, 1976; Wang, Long 2022]

- So what can we do?
  - Bin Yu advocates for the importance of stability

    [Yu, 2013, 2020; Yu, Kumbier 2020]
  - Microcredit: multiple randomized controlled trials in different countries by different researchers
  - Importance of incentivizing follow-up work, replication
  - Not always feasible to convene multiple teams
  - Explainability as a form of stability

    [Arrieta et al 2020; Doshi-Velez et al 2017; Zhang et al 2020; Mittelstadt et al 2019]
    - **Best practice:** Visualizing and investigating the data

A fast, easy-to-use check for stability

- Might worry about stability/generalization if could drop a tiny fraction of data and change substantive conclusions

A fast, easy-to-use check for stability

- Might worry about stability/generalization if could drop a tiny fraction of data and change substantive conclusions
  - E.g. in a study of microcredit with ~16,500 data points, we can drop one data point to flip the sign of the effect

A fast, easy-to-use check for stability

- Might worry about stability/generalization if could drop a tiny fraction of data and change substantive conclusions
  - E.g. in a study of microcredit with ~16,500 data points, we can drop one data point to flip the sign of the effect
    - We can drop 15 data points to get a statistically significant effect of the opposite sign

A fast, easy-to-use check for stability

- Might worry about stability/generalization if could drop a tiny fraction of data and change substantive conclusions
  - E.g. in a study of microcredit with ~16,500 data points, we can drop one data point to flip the sign of the effect
    - We can drop 15 data points to get a statistically significant effect of the opposite sign
- **Challenge**: Way too costly to check every data subset

A fast, easy-to-use check for stability

- Might worry about stability/generalization if could drop a tiny fraction of data and change substantive conclusions
  - E.g. in a study of microcredit with ~16,500 data points, we can drop one data point to flip the sign of the effect
    - We can drop 15 data points to get a statistically significant effect of the opposite sign
- **Challenge**: Way too costly to check every data subset
  - Would take $>10^{44}$ years to check the data analysis above

A fast, easy-to-use check for stability

- Might worry about stability/generalization if could drop a tiny fraction of data and change substantive conclusions
  - E.g. in a study of microcredit with ~16,500 data points, we can drop one data point to flip the sign of the effect
    - We can drop 15 data points to get a statistically significant effect of the opposite sign
- **Challenge**: Way too costly to check every data subset
  - Would take >10^{44} years to check the data analysis above
- **We show**: an *approximation* is fast, easy-to-use, accurate

A fast, easy-to-use check for stability

- Might worry about stability/generalization if could drop a tiny fraction of data and change substantive conclusions
    - E.g. in a study of microcredit with ~16,500 data points, we can drop one data point to flip the sign of the effect
        - We can drop 15 data points to get a statistically significant effect of the opposite sign
- **Challenge**: Way too costly to check every data subset
    - Would take $>10^{44}$ years to check the data analysis above
- **We show**: an *approximation* is fast, easy-to-use, accurate
    - Fast: seconds to run on data analysis above

A fast, easy-to-use check for stability

- Might worry about stability/generalization if could drop a tiny fraction of data and change substantive conclusions
  - E.g. in a study of microcredit with ~16,500 data points, we can drop one data point to flip the sign of the effect
    - We can drop 15 data points to get a statistically significant effect of the opposite sign
- **Challenge**: Way too costly to check every data subset
  - Would take $>10^{44}$ years to check the data analysis above
- **We show**: an *approximation* is fast, easy-to-use, accurate
  - Fast: seconds to run on data analysis above
  - Easy-to-use: no need for user to derive equations

A fast, easy-to-use check for stability

- Might worry about stability/generalization if could drop a tiny fraction of data and change substantive conclusions
  - E.g. in a study of microcredit with ~16,500 data points, we can drop one data point to flip the sign of the effect
    - We can drop 15 data points to get a statistically significant effect of the opposite sign
- **Challenge**: Way too costly to check every data subset
  - Would take >10^{44} years to check the data analysis above
- **We show**: an *approximation* is fast, easy-to-use, accurate
  - Fast: seconds to run on data analysis above
  - Easy-to-use: no need for user to derive equations
  - Accurate: We have theory. But more importantly, we return the dropped points, so can check directly

A fast, easy-to-use check for stability

- Might worry about stability/generalization if could drop a tiny fraction of data and change substantive conclusions
  - E.g. in a study of microcredit with ~16,500 data points, we can drop one data point to flip the sign of the effect
    - We can drop 15 data points to get a statistically significant effect of the opposite sign
- **Challenge**: Way too costly to check every data subset
  - Would take $>10^{44}$ years to check the data analysis above
- **We show**: an *approximation* is fast, easy-to-use, accurate
  - Fast: seconds to run on data analysis above
  - Easy-to-use: no need for user to derive equations
  - Accurate: We have theory. But more importantly, we return the dropped points, so can check directly
- Note: any useful data analysis is sensitive to *some* change

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It’s not just non-significance**

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It’s not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)
  - $p < 0.01$ for a positive effect of lottery win on health

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)
  - $p < 0.01$ for a positive effect of lottery win on health
    - We find: dropping 10 data points (0.05%) changes significance

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It’s not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)
  - $p < 0.01$ for a positive effect of lottery win on health
    - We find: dropping 10 data points (0.05%) changes significance
- **Using Bayes or more-complex models isn’t a panacea**

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)
  - $p < 0.01$ for a positive effect of lottery win on health
    - We find: dropping 10 data points (0.05%) changes significance
- **Using Bayes or more-complex models isn't a panacea**
  - Meager 2022: aggregate Bayesian analysis across 7 randomized controlled trials of microcredit

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)
  - $p < 0.01$ for a positive effect of lottery win on health
    - We find: dropping 10 data points (0.05%) changes significance
- **Using Bayes or more-complex models isn't a panacea**
  - Meager 2022: aggregate Bayesian analysis across 7 randomized controlled trials of microcredit
  - Carefully chosen models

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)
  - $p < 0.01$ for a positive effect of lottery win on health
    - We find: dropping 10 data points (0.05%) changes significance
- **Using Bayes or more-complex models isn't a panacea**
  - Meager 2022: aggregate Bayesian analysis across 7 randomized controlled trials of microcredit
  - Carefully chosen models
  - Can drop <0.03% of data to change a Bayesian version of statistical significance

What makes an analysis non-robust?

- Upcoming real data analyses: *all awesomely reproducible!*
- **It's not just non-significance**
  - Lottery in Oregon, USA
    - Winners could sign up for Medicaid (healthcare)
  - Finkelstein et al 2012, >21,000 data points (surveys)
  - $p < 0.01$ for a positive effect of lottery win on health
    - We find: dropping 10 data points (0.05%) changes significance
- **Using Bayes or more-complex models isn't a panacea**
  - Meager 2022: aggregate Bayesian analysis across 7 randomized controlled trials of microcredit
  - Carefully chosen models
  - Can drop <0.03% of data to change a Bayesian version of statistical significance
  - Can drop <0.1% of data to change the effect sign

What makes an analysis non-robust?

- **It’s not just that everything is non-robust**

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
    1. Both our theory and simulations show that dropping-data sensitivity reflects a signal-to-noise ratio

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  1. Both our theory and simulations show that dropping-data sensitivity reflects a signal-to-noise ratio
  2. Effect of cash transfers on food consumption in poor households
     - Angelucci & De Giorgi 2009, >10,000 data points

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  1. Both our theory and simulations show that dropping-data sensitivity reflects a signal-to-noise ratio
  2. Effect of cash transfers on food consumption in poor households
     - Angelucci & De Giorgi 2009, >10,000 data points
     - We have to drop >4% data to change conclusions

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  1. Both our theory and simulations show that dropping-data sensitivity reflects a signal-to-noise ratio
  2. Effect of cash transfers on food consumption in poor households
     - Angelucci & De Giorgi 2009, >10,000 data points
     - We have to drop >4% data to change conclusions
- **Removing outliers isn't a panacea**

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  1. Both our theory and simulations show that dropping-data sensitivity reflects a signal-to-noise ratio
  2. Effect of cash transfers on food consumption in poor households
     - Angelucci & De Giorgi 2009, >10,000 data points
     - We have to drop >4% data to change conclusions
- **Removing outliers isn't a panacea**
  - Angelucci & De Giorgi 2009 look at "spillover" effect on non-poor households in the same village

What makes an analysis non-robust?

- **It’s not just that everything is non-robust**
  1. Both our theory and simulations show that dropping-data sensitivity reflects a signal-to-noise ratio
  2. Effect of cash transfers on food consumption in poor households
     - Angelucci & De Giorgi 2009, >10,000 data points
     - We have to drop >4% data to change conclusions
- **Removing outliers isn’t a panacea**
  - Angelucci & De Giorgi 2009 look at “spillover” effect on non-poor households in the same village
  - Original analysis removes the largest responses

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  1. Both our theory and simulations show that dropping-data sensitivity reflects a signal-to-noise ratio
  2. Effect of cash transfers on food consumption in poor households
     - Angelucci & De Giorgi 2009, >10,000 data points
     - We have to drop >4% data to change conclusions
- **Removing outliers isn't a panacea**
  - Angelucci & De Giorgi 2009 look at "spillover" effect on non-poor households in the same village
  - Original analysis removes the largest responses
  - We can drop 3 points of >4,000 & change significance

What makes an analysis non-robust?

- **It's not just that everything is non-robust**
  1. Both our theory and simulations show that dropping-data sensitivity reflects a signal-to-noise ratio
  2. Effect of cash transfers on food consumption in poor households
     - Angelucci & De Giorgi 2009, >10,000 data points
     - We have to drop >4% data to change conclusions
- **Removing outliers isn't a panacea**
  - Angelucci & De Giorgi 2009 look at "spillover" effect on non-poor households in the same village
  - Original analysis removes the largest responses
  - We can drop 3 points of >4,000 & change significance
- **p-hacking isn't robust to dropping a small data fraction:** • michaelwiebe.com/blog/2021/01/amip
  - rgiordan.github.io/robustness/2021/09/17/amip_p_hacking.html

Conclusions & Resources

- We review some challenges and mitigations in trusting data analyses (even when everyone is well-meaning)
- **Paper**: Broderick, Gelman, Meager, Smith, Zheng. “Toward a taxonomy of trust for probabilistic machine learning.” *Science Advances*, 2023.

- We present a way to check if there exists a very small fraction of data you can drop to change decisions
- **Paper**: Giordano*, Meager*, Broderick “An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?” ArXiv: 2011.14999
- **Code, etc**: `github.com/rgiordan/zaminfluence`
- **Biology**: Shiffman, Giordano, Broderick “Could dropping a few cells change the takeaways from differential expression?” ArXiv.

Additional References (1/5)

Abeysooriya et al. Gene name errors: Lessons not learned. *PLOS Computational Biology*, 2021.

Angelucci & De Giorgi. Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? *American Economic Review*, 2009.

Angelucci et al. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 2015.

Arrieta et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 2020.

Baggerly & Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics*, 2009.

Begley & Ellis. Raise standards for preclinical cancer research. *Nature*, 2012.

Box, Science and statistics. *Journal of the American Statistical Association*, 1976.

Charles et al. Iron-deficiency anaemia in rural Cambodia: community trial of a novel iron supplementation technique. *European Journal of Public Health*, 2011.

Crépon et al. Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment. *The Quarterly Journal of Economics*, 2013.

Additional References (2/5)

Doshi-Velez et al. Accountability of AI under the law: The role of explanation. arXiv: 1711.01134, 2017.

Federal Reserve Board's Division of Research and Statistics, Changes in U.S. Family Finances from 2019 to 2022: Evidence from the Survey of Consumer Finances. October 2023. (p. 11)

Finkelstein et al. The Oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics*, 2012.

Gebru et al. Datasheets for datasets. *Communications of the ACM*, 2021.

Haibe-Kains et al. Transparency and reproducibility in artificial intelligence. *Nature*, 2020.

Heaven. Hundreds of AI tools have been built to catch covid. None of them helped. *MIT Technology Review*, 2021.

Heil et al. Reproducibility standards for machine learning in the life sciences. *Nature Methods*, 2021.

Herndon et al. Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 2014.

Lewis. Autocorrect errors in Excel still creating genomics headache. *Nature: News*, 2021.

McKinney et al. International evaluation of an AI system for breast cancer screening. *Nature*, 2020.

Additional References (3/5)

Meager. Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature. *American Economic Review*, 2022.

Mittelstadt et al. Explaining explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

NASA Earth Observatory, Research satellites for atmospheric sciences, 1978–present: Serendipity and stratospheric ozone, 2001. Accessed 2021.

Open Science Collaboration, Estimating the reproducibility of psychological science. *Science*, 2015.

Pukelsheim, Robustness of Statistical Gossip and the Antarctic Ozone Hole. *Letters to the Editor: The IMS Bulletin*, 1990.

Rappaport et al. Randomized controlled trial assessing the efficacy of a reusable fish-shaped iron ingot to increase hemoglobin concentration in anemic, rural Cambodian women. *The American Journal of Clinical Nutrition*, 2017.

Reinhart & Rogoff, Growth in a time of debt. *The American Economic Review*, 2010.

Roberts et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 2021.

Additional References (4/5)

Silberzahn et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 2018.

Stodden et al. *Implementing Reproducible Research*, CRC Press, 2014.

Stodden et al. Enhancing reproducibility for computational methods. *Science*, 2016.

Vartak et al. MODELDB: A system for machine learning model management. *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, 2016.

Wang & Long. Addressing Common Misuses and Pitfalls of *P* values in Biomedical Research. *Cancer Research*, 2022.

Wieringa et al. Low Prevalence of Iron and Vitamin A Deficiency among Cambodian Women of Reproductive Age. *Nutrients*, 2016.

Winkler et al. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, 2019.

Yu. Stability. *Bernoulli*, 2013.

Yu. Stability expanded in reality. *Harvard Data Science Review*, 2020.

Yu, Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 2020.

Additional References (5/5)

Zeeberg et al. Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*, 2004.

Zhang et al. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making, *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

Ziemann et al. Gene name errors are widespread in the scientific literature. *Genome Biology*, 2016.

Image References

https://commons.wikimedia.org/wiki/File:Money_saving_growth.jpg (Creative Commons CC0 1.0 Universal Public Domain Dedication)

https://commons.wikimedia.org/wiki/File:Chest_X-ray_2346.jpg (Creative Commons CC0 1.0 Universal Public Domain Dedication)

https://commons.wikimedia.org/wiki/File:Wikimedia_in_Education_illustration_books.svg (Creative Commons CC0 1.0 Universal Public Domain Dedication)